

1. Fundamente

Operatiile aritmetice fundamentale : adunarea si inmultirea.
Sunt necesare pentru a evalua un polinom $P(x)$ intr-un punct x .

1.1. Evaluarea unui polinom: $P(x) = 2x^4 + 3x^3 - 3x^2 + 5x - 1$, $x = \frac{1}{2}$

I. Metoda directă

$$P\left(\frac{1}{2}\right) = 2 * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} + 3 * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} - 3 * \frac{1}{2} * \frac{1}{2} + 5 * \frac{1}{2} - 1 = \frac{5}{4}$$

Numarul de înmultiri necesare este 10, la care se adauga 4 adunari

II. Găsim puterile numărului $\frac{1}{2}$

$$\frac{1}{2} * \frac{1}{2} = \left(\frac{1}{2}\right)^2$$

$$\left(\frac{1}{2}\right)^2 * \frac{1}{2} = \left(\frac{1}{2}\right)^3$$

$$\left(\frac{1}{2}\right)^3 * \frac{1}{2} = \left(\frac{1}{2}\right)^4$$

- acum, putem face o simplă însumare a termenilor:

$$P\left(\frac{1}{2}\right) = 2 * \left(\frac{1}{2}\right)^4 + 3 * \left(\frac{1}{2}\right)^3 - 3 * \left(\frac{1}{2}\right)^2 + 5 * \frac{1}{2} - 1 = \frac{5}{4}$$

- avem 3 înmulțiri ale lui $1/2$, plus alte 4 înmulțiri
- am redus numărul total de înmulțiri la 7, cu aceleași 4 adunări
- este reducerea de la 14 la 11 operații o îmbunătățire semnificativă?
- dacă polinomul trebuie evaluat pentru diferite valori ale lui x , de mai multe ori pe secundă, atunci diferența poate fi esențială

III. Înmultirea imbricată (metoda lui Horner)

$$\begin{aligned} P(x) &= -1 + x(5 - 3x + 3x^2 + 2x^3) = \\ &= -1 + x(5 + x(-3 + 3x + 2x^2)) = \\ &= -1 + x(5 + x(-3 + x(3 + 2x))) \end{aligned}$$

Evaluarea se face dinspre interior spre exterior.

înmulțim $\frac{1}{2} * 2$,	adunăm $+3 \rightarrow 4$
înmulțim $\frac{1}{2} * 4$,	adunăm $-3 \rightarrow -1$
înmulțim $\frac{1}{2} * -1$,	adunăm $+5 \rightarrow \frac{9}{2}$
înmulțim $\frac{1}{2} * \frac{9}{2}$,	adunăm $-1 \rightarrow \frac{5}{4}$

In general, un polinom de gradul d poate fi evaluat utilizând aceasta metoda folosind d înmultiri și d adunari.

Forma generală a unui polinom

$$C_1 + C_2 x + C_3 x^2 + C_4 x^3 + C_5 x^4 = C_1 + x(C_2 + x(C_3 + x(C_4 + x(C_5))))$$

Pentru interpolare, vom avea nevoie de forma:

$$c_1 + (x - r_1)(c_2 + (x - r_2)(c_3 + (x - r_3)(c_4 + (x - r_4)(c_5))))$$

unde r_1, r_2, r_3, r_4 - puncte de bază

Exemplul 1

- găsiți o metodă eficientă pentru evaluarea polinomului
 $P(x) = 4x^5 + 7x^8 - 3x^{11} + 2x^{14}$
- ideea este de a factoriza x^5 din fiecare termen și de a scrie polinomul în x^3 :

$$\begin{aligned} P(x) &= x^5(4 + 7x^3 - 3x^6 + 2x^9) \\ &= x^5 * (4 + x^3 * (7 + x^3 * (-3 + x^3 * (2)))) \end{aligned}$$

- pentru fiecare punct x , trebuie să calculăm mai întâi $x * x = x^2$,
 $x * x^2 = x^3$, și $x^2 * x^3 = x^5$
- aceste trei înmulțiri, împreună cu înmulțirea lui x^5 , și cele trei înmulțiri și trei adunări necesare evaluării polinomului de gradul 3 în x^3 dau numărul total de operații necesar evaluării acestui polinom, și anume 7 înmulțiri și 3 adunări

1.2. Numere binare

Un nr în baza 2: $\dots b_2 \cdot 2^2 + b_1 \cdot 2^1 + b_0 \cdot 2^0 + b_{-1} \cdot 2^{-1} + b_{-2} \cdot 2^{-2} \dots$

$$(4)_{10} = (100)_2$$

$$\left(\frac{3}{4}\right)_{10} = (0,75)_{10} = (0,11)_2$$

1.2.1. Conversia din zecimal în binar

$$(53,7)_{10} = (53)_{10} + (0,7)_{10}$$

Partea întreagă: $(53)_{10} = (110101)_2$

Partea fracționară:

$$0,7 \times 2 \quad 0,4 + 1$$

$$0,4 \times 2 \quad 0,8 + 0$$

$$0,8 \times 2 \quad 0,6 + 1$$

$$0,6 \times 2 \quad 0,2 + 1$$

$$0,2 \times 2 \quad 0,4 + 0$$

$$0,4 \times 2 \quad 0,8 + 0$$

Observăm că acest proces se repetă la infinit după 4 pași.

$$\Rightarrow (0.7)_{10} = (0.1011001100110\dots)_2 = (0.10110)_{\overline{2}}$$

$$\Rightarrow (53.7)_{10} = (110101.10110)_{\overline{2}}$$

1.2.2. Conversia din binar în zecimal

$$(10101)_2 = 1 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = (21)_{10}$$

$$(0.1011)_2 = 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2^2} + 1 \cdot \frac{1}{2^3} + 1 \cdot \frac{1}{2^4} = \left(\frac{11}{16}\right)_{10}$$

$$\circ (0.1011)_{\overline{2}}$$

Înmulțim nr. cu $2^4 \Rightarrow$ deplasare la stânga cu 4 poziții

$$2^4 \cdot x = 1011.1011$$

$$x = 0000.1011$$

$$\Rightarrow x(2^4 - 1) = (1011)_2 = (11)_{10}$$

$$\Rightarrow x = (0.1011)_{\overline{2}} = \left(\frac{11}{15}\right)_{10}$$

$$\circ x = (0.10101)_{\overline{2}}$$

$$y = 2^2 \cdot x = (10.101)_{\overline{2}}$$

$$\{y\} = (0.101)_{\overline{2}} = z$$

$$\Rightarrow 2^3 \cdot z = (101.101)_2$$

$$z = (000.101)_2$$

$$\Rightarrow z(2^3 - 1) = (101)_2 = 5 \Rightarrow z = \frac{5}{7}$$

$$\Rightarrow y = (10)_2 + \frac{5}{7} = 2 + \frac{5}{7} = \frac{19}{7}$$

$$\Rightarrow x = 2^{-2} \cdot y = \left(\frac{19}{28}\right)_{10}$$

1.3. Reprezentarea în virgulă flotantă a numerelor reale

1.3.1. Formate de virgulă flotantă

Un număr în virgulă flotantă este format din trei părți:

a) semnul (+ sau -)

b) mantisa (care conține șirul de biți semnificativi)

c) exponentul

Există 3 nivele de precizie:

I. simplă

→ 32

II. dublă

→ 64

III. extinsă (lungă)

→ 80

BITI ALOCAȚI

precizia	semnul	exponentul	mantisa
simplă	1	8	23
dublă	1	11	52
dublă lungă	1	15	64

Definiția 1

Numărul **epsilon mașină**, notat cu ϵ_{mach} , este distanța dintre 1 și cel mai mic număr în virgulă flotantă mai mare decât 1. Pentru standardul IEEE de virgulă flotantă în dublă precizie, avem că

$$\epsilon_{\text{mach}} = 2^{-52}.$$

- numărul zecimal $9.4 = (1001.0110)_2$ este aliniat la stânga astfel:

$$+1.\boxed{0010110011001100110011001100110011001100110011001100110011001100}110 \dots \times 2^3,$$

unde am pus într-un chenar primii 52 de biți ai mantisei

1.3.1 Formate de virgulă flotantă

Algoritmul 1 (Regula IEEE a rotunjirii la cea mai apropiată valoare)

Pentru dubla precizie, dacă al 53-lea bit de la dreapta virgulei binare este 0, atunci **rotunjim** în jos (trunchiem după bitul 52). Dacă bitul al 53-lea este 1, atunci rotunjim în sus (adunăm 1 la bitul 52), cu excepția cazului în care biții de după 1 sunt 0, caz în care 1 este adunat la bitul 52 dacă și numai dacă bitul 52 este 1.

- pentru numărul 9.4 discutat anterior, cel de-al 53-lea bit din dreapta virgulei binare este 1 și este urmat de alți biți care nu sunt toți zero
- conform regulii rotunjirii la cea mai apropiată valoare, trebuie să facem o rotunjire în sus, adică să adunăm un 1 la bitul 52
- prin urmare, numărul în virgulă flotantă care îl reprezintă pe 9.4 este

$$+1.\boxed{0010110011001100110011001100110011001100110011001100110011001101} \times 2^3.$$

(7)

Definiția 2

Notăm numărul în virgulă flotantă în dublă precizie IEEE asociat lui x , folosind regula rotunjirii la cea mai apropiată valoare, cu $\text{fl}(x)$.

- în operațiile aritmetice realizate în calculator, numărul real x este înlocuit cu șirul de biți $\text{fl}(x)$
- conform definiției de mai sus, $\text{fl}(9.4)$ este numărul în reprezentare binară dat de (7)
- am ajuns la această reprezentare în virgulă flotantă înlăturând partea infinită dată de $0.1100 \times 2^{-52} \times 2^3 = 0.0110 \times 2^{-51} \times 2^3 = 0.4 \times 2^{-48}$ din capătul din dreapta al numărului, și apoi adunând $2^{-52} \times 2^3 = 2^{-49}$ în pasul de rotunjire
- prin urmare,

$$\begin{aligned}\text{fl}(9.4) &= 9.4 + 2^{-49} - 0.4 \times 2^{-48} \\ &= 9.4 + (1 - 0.8)2^{-49} \\ &= 9.4 + 0.2 \times 2^{-49}.\end{aligned}\tag{8}$$

- cu alte cuvinte, un calculator care folosește reprezentarea în dublă precizie și regula rotunjirii la cea mai apropiată valoare face o eroare de 0.2×2^{-49} atunci când stochează numărul 9.4
- vom spune că 0.2×2^{-49} este **eroarea de rotunjire**
- ceea ce trebuie reținut este faptul că numărul în virgulă flotantă care îl reprezintă pe 9.4 nu este egal cu 9.4, deși este foarte aproape de această valoare
- pentru a cuantifica această apropiere, vom folosi definiția standard a erorii

Definiția 3

Fie x_c o versiune calculată a valorii exacte x . Atunci

$$\text{eroarea absolută} = |x_c - x|,$$

și

$$\text{eroarea relativă} = \frac{|x_c - x|}{|x|},$$

dacă această din urmă cantitate există ($x \neq 0$).

Algoritmul 2 (Eroarea relativă de rotunjire)

În cadrul standardului IEEE, eroarea relativă de rotunjire $\text{fl}(x)$ este mai mică decât jumătate din numărul epsilon mașină:

$$\frac{|\text{fl}(x) - x|}{|x|} \leq \frac{1}{2} \epsilon_{\text{mach}}.\tag{9}$$

- în cazul numărului $x = 9.4$, am găsit eroarea de rotunjire în (8), care trebuie să satisfacă (9):

$$\frac{|\text{fl}(9.4) - 9.4|}{9.4} = \frac{0.2 \times 2^{-49}}{9.4} = \frac{8}{47} \times 2^{-52} < \frac{1}{2} \epsilon_{\text{mach}}.$$

Exemplul 2

- găsiți reprezentarea în dublă precizie $fl(x)$ și eroarea de rotunjire pentru $x = 0.4$
- deoarece $(0.4)_{10} = (0.\overline{0110})_2$, alinierea la stânga a acestui număr binar ne dă:

$$\begin{aligned}0.4 &= 1.\overline{100110} \times 2^{-2} \\&= +1.\boxed{1001100110011001100110011001100110011001100110011001} \\&\quad 100110\dots \times 2^{-2}.\end{aligned}$$

- prin urmare, în conformitate cu regula de rotunjire, $fl(0.4)$ este $+1.\overline{10011001100110011001100110011001100110011010} \times 2^{-2}$.
- aici, 1 a fost adunat la bitul 52, ceea ce a determinat o schimbare și a bitului 51, ca urmare a transportului din adunarea binară
- analizând cu atenție, am înlăturat $2^{-53} \times 2^{-2} + 0.\overline{0110} \times 2^{-54} \times 2^{-2}$ în cadrul trunchierii și am adunat $2^{-52} \times 2^{-2}$ prin rotunjirea în sus

Prof. Dr. Călin-Adrian POPA

Matematici Asistate de Calculator

1.3.1 Formate de virgulă flotantă

- prin urmare,

$$\begin{aligned}\text{fl}(0.4) &= 0.4 - 2^{-55} - 0.4 \times 2^{-56} + 2^{-54} \\ &= 0.4 + 2^{-54}(-1/2 - 0.1 + 1) \\ &= 0.4 + 2^{-54}(0.4) \\ &= 0.4 + 0.1 \times 2^{-52}.\end{aligned}$$

- observăm că eroarea relativă de rotunjire pentru 0.4 este $0.1/0.4 \times \epsilon_{\text{mach}} = 1/4 \times \epsilon_{\text{mach}}$, conform cu (9)

1.3.2. Reprezentarea numerelor în calculator

fiecărui număr în virgulă flotantă în dublă precizie îi este asignat un cuvânt de 8 octeți (1 octet = 8 biți), sau 64 de biți, pentru a stoca cele trei părți ale sale
fiecare astfel de cuvânt are forma

$$\boxed{se_1 e_2 \dots e_{11} b_1 b_2 \dots b_{52}}, \quad (10)$$

unde mai întâi este stocat semnul, urmat de 11 biți reprezentând exponentul și de 52 de biți care urmează virgulei zecimale, care reprezintă mantisa

bitul de semn s este 0 pentru un număr pozitiv și 1 pentru un număr negativ

cei 11 biți reprezentând exponentul vin din numărul binar pozitiv care rezultă din adunarea lui $2^{10} - 1 = 1023$ la exponent, cel puțin pentru exponenți între -1022 și 1023

- exponentul special 0, adică $e_1 e_2 \dots e_{11} = (000\ 0000\ 0000)_2$, denotă de asemenea o abatere de la forma standard a unui număr în virgulă flotantă
- în acest caz, numărul este interpretat ca numărul în virgulă flotantă nenormalizat

- adică, doar în acest caz, bitul cel mai din stânga nu este presupus a fi egal cu 1
- aceste numere nenormalizate se numesc numere în virgulă flotantă **subnormale**
- ele extind gama numerelor foarte mici cu încă câteva ordine de mărime
- prin urmare, $2^{-52} \times 2^{-1022} = 2^{-1074}$ este cel mai mic număr diferit de zero care este reprezentabil în dublă precizie
- cuvântul corespunzător care va fi stocat în calculator este

- trebuie făcută distincția între cel mai mic număr reprezentabil 2^{-1074} și $\epsilon_{\text{mach}} = 2^{-52}$
- sunt multe numere mai mici decât ϵ_{mach} care sunt reprezentabile, chiar dacă adunându-le la 1 nu vom obține niciun efect
- pe de altă parte, numerele în dublă precizie mai mici decât 2^{-1074} nu pot fi reprezentate deloc
- numerele subnormale includ și cel mai important număr, 0
- de fapt, reprezentarea subnormală include două numere în virgulă flotantă diferite, $+0$ și -0 , care sunt tratate în calcule ca fiind același număr real
- reprezentarea în virgulă flotantă a lui $+0$ are bitul de semn $s = 0$, biții de exponent $e_1 \dots e_{11} = 00000000000$, și mantisa 52 de zerouri; pe scurt, toți cei 64 de biți sunt zero
- formatul hexazecimal pentru $+0$ este 0000000000000000
- pentru numărul -0 , totul este exact la fel, cu excepția bitului de semn $s = 1$
- formatul hexazecimal pentru -0 este 8000000000000000

- adunarea în calculator constă din alinierea virgulelor zecimale ale celor două numere de adunat, adunarea lor, și stocarea rezultatului tot ca un număr în virgulă flotantă
- adunarea propriu-zisă poate fi realizată cu precizie mai mare (cu mai mult de 52 de biți) deoarece are loc într-un registru dedicat special pentru acest scop
- după adunare, rezultatul trebuie rotunjit înapoi la 52 de biți după virgula binară pentru a fi stocat ca un număr în virgulă flotantă
- de exemplu, adunarea lui 1 la 2^{-53} va apărea după cum urmează:

- acest număr este stocat ca $1.0 \times 2^0 = 1$, conform regulii de rotunjire
 - prin urmare, $1 + 2^{-53}$ este egal cu 1 în aritmetica în dublă precizie
- IEEE

Exemplul 4

- găsiți suma în virgulă flotantă în dublă precizie $(1 + 3 \times 2^{-53}) - 1$
- bineînțeles, în aritmetica reală, răspunsul este 3×2^{-53}
- însă calculele în aritmetica în virgulă flotantă pot da un rezultat diferit
- observăm că $3 \times 2^{-53} = 2^{-52} + 2^{-53}$
- prima adunare este

[illegible]

- acesta este cazul de excepție pentru regula rotunjirii
- deoarece bitul 52 din sumă este 1, trebuie să facem o rotunjire în sus, ceea ce înseamnă să adunăm 1 la bitul 52
- după transport, obținem

[illegible]

care este reprezentarea lui $1 + 2^{-51}$

- prin urmare, după ce scădem 1, rezultatul va fi 2^{-51} , care este egal cu $2^{\epsilon_{\text{mach}}} = 4 \times 2^{-53}$
- din nou, se poate observa diferența dintre aritmetica în calculator și aritmetica exactă

2. Rezolvarea ecuatiilor

Def 1: Funcția $f(x)$ are o rădăcină în $x=r$ dacă $f(r)=0$

Def 2: r este un punct fix al funcției g dacă $g(r) = r$

Proprietate: Orice ecuație $f(x) = 0$ poate fi transformată într-o problemă de punct fix.

$$\text{Def: } x^3 + x - 1 = 0$$

$$1) \quad x^3 + x - 1 = 0 \quad | + 1 - x \Rightarrow \quad x = 1 - x^3 = g(x)$$

$$2) \quad x^3 + x - 1 = 0 \quad | +1 - x \quad \Rightarrow \quad x^3 = 1 - x \quad \Rightarrow \quad x = \sqrt[3]{1-x} = g(x)$$

$$\begin{aligned} \text{b)} \quad x^3 + x - 1 &= 0 \quad | +2x^3 + 1 \Rightarrow 3x^3 + x = 2x^3 + 1 \\ &\Rightarrow x(3x^2 + 1) = 2x^3 + 1 \\ &\Rightarrow x = \frac{2x^3 + 1}{3x^2 + 1} = g(x) \end{aligned}$$

2. 1. Metoda bisectiei $\rightarrow f(x)=0$

2.1.1. Găsirea limitelor între care se află o rădăcină

\rightarrow verificăm dacă \exists o rădăcină

$$\left. \begin{array}{l} [a, b] \quad f(a) \cdot f(b) < 0 \\ f \text{ - continuă și } r \in [a, b] \end{array} \right\} \Rightarrow f(r) = 0$$

Teorema 1 (Teorema valorii intermediare)

Fie f o funcție continuă pe intervalul $[a, b]$. Atunci f ia orice valoare între $f(a)$ și $f(b)$. Mai precis, dacă y este un număr între $f(a)$ și $f(b)$, atunci există un număr c care satisface $a \leq c \leq b$ astfel încât $f(c) = y$.

Teorema 2

Fie f o funcție continuă pe $[a, b]$, care satisface $f(a)f(b) < 0$. Atunci f are o rădăcină între a și b , adică există un număr r care satisface $a < r < b$ și $f(r) = 0$.

- în Figura 1, $f(0)f(1) = (-1)(1) < 0$
- există o rădăcină imediat la stânga lui 0.7
- cum putem rafina presupunerea inițială despre locația rădăcinii cu mai multe zecimale exacte?

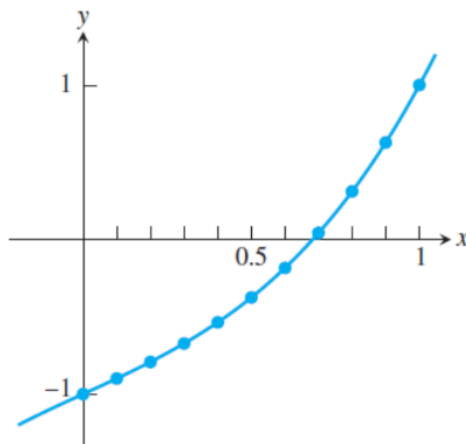


Figura 1: Graficul funcției $f(x) = x^3 + x - 1$. Funcția are o rădăcină între 0.6 și 0.7.

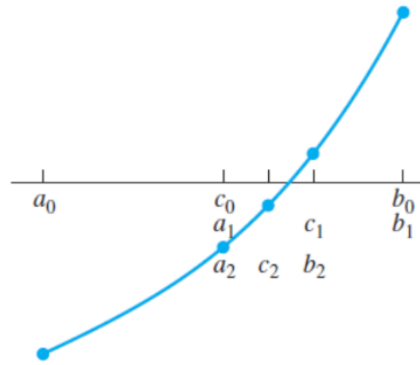


Figura 2: Metoda biseției. În primul pas, semnul lui $f(c_0)$ este verificat. Deoarece $f(c_0)f(b_0) < 0$, asignăm $a_1 = c_0$, $b_1 = b_0$, și intervalul este înlocuit cu jumătatea lui dreaptă $[a_1, b_1]$. În al doilea pas, subintervalul este înlocuit cu jumătatea lui stângă $[a_2, b_2]$.

Algoritmul 3 (Metoda biseției)

Dându-se un interval inițial $[a, b]$ astfel încât $f(a)f(b) < 0$

```
while  $(b - a)/2 > \text{TOL}$ 
     $c = (a + b)/2$ 
    if  $f(c) = 0$ , stop, end
    if  $f(a)f(c) < 0$ 
         $b = c$ 
    else
         $a = c$ 
    end
```

end

Intervalul final $[a, b]$ conține o rădăcină.

Aproximarea rădăcinii este $(a + b)/2$.

```
function x = metoda_bisectiei(f,a, b, tol)

if f(a) * f(b) >= 0
    error('Conditia f(a) * f(b) < 0 nu este satisfacuta!')
end

while (b - a) / 2 > tol
    c = (a + b) / 2;
    if f(c) == 0
        break;
    end
    if f(a) * f(c) < 0
        b = c;
    else
        a = c;
    end
end

x = (a + b) / 2;
```